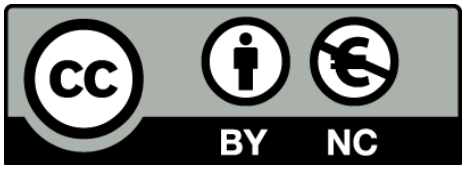
Estimadores núcleos y polinomios locales.

Francisco Parra Rodriguez

Doctor en Ciencias Económicas. UNED.



Modelos de regresión no paramétricos

Los modelos de regresión paramétricos suponen que los datos observados provienen de variables aleatorias cuya distribución es conocida, salvo por la presencia de algunos parámetros cuyo valor se desconoce.
$y = \beta_0 + \beta_1 x + \varepsilon$, con $\varepsilon \approx N(0, \sigma^2)$

Este es un modelo estadístico con tres parámetros desconocidos: β_0; β_1 y σ^2.
Una formulación general de un modelo de regresión paramétrico es la siguiente:

$y = m(x_i;\theta) + \varepsilon_i$, $i = 1,...,n$, $\theta \in \Theta \subseteq \Re^p$
Donde $m(x_i;\theta)$ es una función conocida de x y de θ, que es desconocido, $\varepsilon_1 ... \varepsilon_n$ es una variable aleatoria idénticamente distribuida con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$. El modelo de regresión lineal simple sería un caso particular con
$\theta = (\beta_o, \beta_1)$ y $m(x_i;\beta_o,\beta_1) = \beta_o + \beta_1 x$.
Se dice que se ajusta el modelo paramétrico cuando se estiman sus parámetros a partir de un conjunto de observaciones que siguen dicho modelo. Por ejemplo, puede usarse el método de mínimos cuadrados o el de máxima verosimilitud, de manera que pueden hacerse predicciones de nuevos valores de y conocido el valor de x, y tener información precisa acerca de la incertidumbre asociada a la estimación y a la predicción.
Estas son algunas de las buenas propiedades de los modelos paramétricos.
Además, en muchas ocasiones los parámetros tienen una interpretación intuitiva en términos relacionados con el problema en estudio.
Sin embargo, si el modelo paramétrico no es adecuado puede ser peor tenerlo ajustado que no tener nada, porque el modelo paramétrico conlleva un grado de exactitud en las afirmaciones que de el se derivan que son adecuadas cuando el modelo es correcto, pero que en caso contrario pueden estar muy alejadas de la realidad.
Los modelos paramétricos presentan un problema fundamental: su estructura es tan rígida que no pueden adaptarse a muchos conjuntos de datos.

En este capítulo presentaremos una alternativa no paramétrica a los modelos de regresión paramétricos usuales.

Se supone que se observa n pares de datos (x_i, y_i) que provienen del siguiente modelo de regresión no paramétrico:

$y_i = m(x_i) + \varepsilon_i$

Donde $\varepsilon_1 ... \varepsilon_n$ es una variable aleatoria idénticamente distribuida con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$, y los valores de la variable explicativa $x_1 ... x_n$ son conocidos, por lo que se dice que el modelo tiene diseño fijo, y dado que la varianza de los errores es constante el modelo es Homocedástico[1].

Considerando (X, Y) una variable aleatoria bivariante con densidad conjunta $f(x, y)$, cabe definir la función de regresión como $m(x) = E(Y / X = x)$, es decir el valor esperado de Y cuando X toma el valor conocido x. Entonces $E(Y / X) = m(X)$, y definiendo $\varepsilon = Y - m(X)$, se tiene que:

$Y = m(X) + \varepsilon$, $E(\varepsilon / X) = 0$, $V(\varepsilon / X) = \sigma^2$

Sean (X_i, Y_i), i=1…n, una muestra aleatoria simple de (X, Y). Estos datos siguen el modelo de regresión no paramétrico:

$Y_i = m(X_i) + \varepsilon_i$, i=1…n.

Una vez establecido el modelo, el paso siguiente consiste en estimarlo (o ajustarlo) a partir de las n observaciones disponibles. Es decir hay que construir un estimador $\hat{m}(x)$ de la función de regresión y un estimador $\hat{\sigma}^2$ de la varianza del error. Los procedimientos de estimación de $m(x)$ se conocen como métodos de suavizado.

El abanico de técnicas disponibles para estimar no paramétricamente la función de regresión es amplísimo e incluye, entre otras, las siguientes:

- Ajuste local de modelos paramétricos. Se basa en hacer varios (o incluso infinitos, desde un punto de vista teórico) ajustes paramétricos teniendo en cuenta únicamente los datos cercanos al punto donde se desea estimar la función.
- Métodos basados en series ortogonales de funciones. Se elige una base ortonormal del espacio vectorial de funciones y se estiman los coeficientes del desarrollo en esa base de la función de regresión. Los ajustes por series de Fourier y mediante wavelets son los dos enfoques más utilizados.
- Suavizado mediante splines. Se plantea el problema de buscar la función $\hat{m}(x)$ que minimiza la suma de los cuadrados de los errores ($e_i = y_i - \hat{m}(x_i)$) más un término que penaliza la falta de suavidad de las funciones $\hat{m}(x)$) candidatas (en términos de la integral del cuadrado de su derivada segunda).
- Técnicas de aprendizaje supervisado. Las redes neuronales, los k vecinos más cercanos y los árboles de regresión se usan habitualmente para estimar $m(x)$.

[1] Si se supone que la varianza es función de la variable explicativa x: $V(\varepsilon_i) = \sigma^2(x_i)$, el modelo sería Heterocedástico.

Función núcleo

Los histogramas son siempre, por naturaleza, funciones discontinuas; sin embargo, en muchos casos es razonable suponer que la función e densidad de la variable que se está estimando es continua. En este sentido, los histogramas son estimadores insatisfactorios. Los histogramas tampoco son adecuados para estimar las modas, a lo sumo, pueden proporcionar “intervalos modales", y al ser funciones constantes a trozos, su primera derivada es cero en casi todo punto, lo que les hace completamente inadecuados para estimar la derivada de la función de densidad.

Los estimadores de tipo núcleo (o kernel) fueron diseñados para superar estas dificultades. La idea original es bastante antigua y se remonta a los trabajos de Rosenblatt y Parzen en los años 50 y primeros 60. Los estimadores kernel son, sin duda, los más utilizados y mejor estudiados en la teoría no paramétrica.

Dada una m.a.s. $X_1 \ldots X_n$ con densidad f, estimamos dicha densidad en un punto t por medio del estimador

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h} \right)$$

donde h es una sucesión de parámetros de suavizado, llamados ventanas o amplitudes de banda (windows, bandwidths) que deben tender a cero ”lentamente" ($h \to 0, nh \to \infty$) para poder asegurar que $\hat{f}$ tiende a la verdadera densidad f de las variables X_i y K es una función que cumple $\int K = 1$. Por ejemplo:

- Núcleo gaussiano:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

- Núcleo Epanechnikov[2]:

$$\frac{3}{4}\left(1 - u^2\right) I_{|u|<1}$$

[2] Otras funciones núcleo que se utilizan para estimar la densidad son:

- Núcleo Triangular:

$$\left(1 - |u|\right) I_{|u|<1}$$

- Núcleo Uniforme:

$$\frac{1}{2} I_{|u|<1}$$

- Núcleo Biweight:

$$\frac{15}{16}\left(1 - u^2\right) I_{|u|<1}$$

- Núcleo Triweight:

$$\frac{35}{32}\left(1 - u^2\right) I_{|u|<1}$$

donde $I_{|u|<1}$ es la función que vale 1 si $|u|<1$ y 0 si $|u|\geq 1$

Para elegir la ventana h podemos seguir la siguiente regla

$$h=\delta_K\left(\frac{3}{8}\right)\pi^{1/10}s_n n^{-1/5}$$

Donde

- n es el tamaño de la muestra
- $s_n=\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^{1/2}$
- δ_K depende del núcleo K, y se calcula como:

$$\delta_K=\left(\frac{\int K^2(t)dt}{\left(\int u^2K(t)dt\right)^2}\right)^{1/5}$$

Por ejemplo:

- Si K es el núcleo gaussiano, entonces $\delta_K=\left(\frac{1}{4\pi}\right)^{1/10}$
- Si K es el núcleo Epanechnikov, entonces $\delta_K=(15)^{1/5}$

Ejemplo

Nuestra muestra $X_1...X_{10}$ es:

2,1	2,6	1,9	4,5	0,7	4,6	5,4	2,9	5,4	0,2

Su desviación típica es $s_n=1{,}8750$, utilizando una función núcleo de Epanechnikov, la ventana h será:

$$h=(15)^{1/5}\times\left(\frac{3}{8}\right)\pi^{1/10}\times 1{,}875\times 10^{-1/5}=0.8550$$

Hacemos una grilla para t que va desde -0,5 a 0,6 con puntos semi-espaciados:

t
-0,5
0,04166667
0,58333333
1,125
1,66666667
2,20833333
2,75
3,29166667
3,83333333
4,375
4,91666667
5,45833333
6

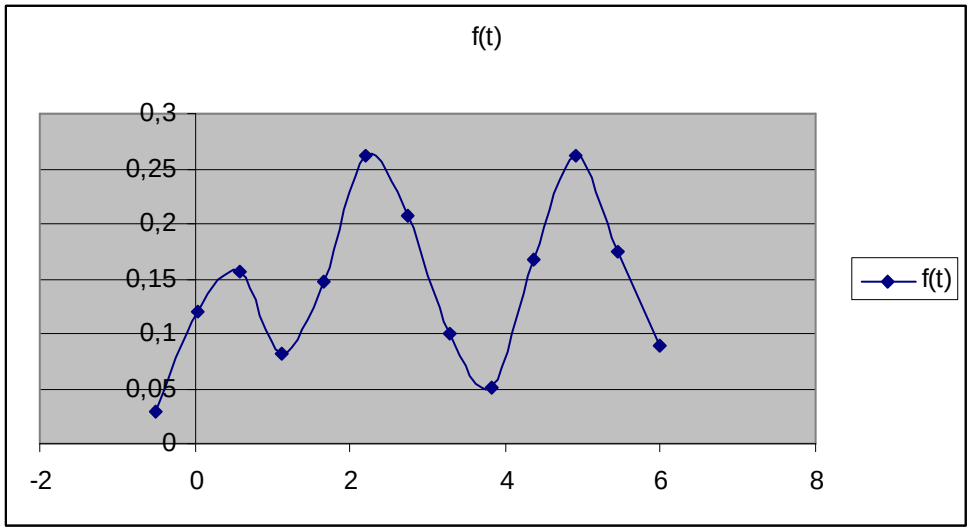
Para cada t_j calculamos $K\left(\frac{t_j - X_i}{h}\right)$:

t	$K\left(\frac{t_j - X_1}{h}\right)$	$K\left(\frac{t_j - X_2}{h}\right)$	$K\left(\frac{t_j - X_3}{h}\right)$	$K\left(\frac{t_j - X_4}{h}\right)$	$K\left(\frac{t_j - X_5}{h}\right)$	$K\left(\frac{t_j - X_6}{h}\right)$	$K\left(\frac{t_j - X_7}{h}\right)$	$K\left(\frac{t_j - X_8}{h}\right)$	$K\left(\frac{t_j - X_9}{h}\right)$	$K\left(\frac{t_j - X_{10}}{h}\right)$	$\sum_{i=1}^{n} K\left(\frac{t_j - X_i}{h}\right)$
-0,5	0	0	0	0	0	0	0	0	0	0,2472864	0,2472864
0,04166667	0	0	0	0	0,3053521	0	0	0	0	0,7242801	1,0296322
0,58333333	0	0	0	0	0,73603573	0	0	0	0	0,59924292	1,33527865
1,125	0	0	0,1337911	0	0,56468848	0	0	0	0	0	0,69847958
1,66666667	0,55735012	0	0,69414293	0	0	0	0	0	0	0	1,25149305
2,20833333	0,73795938	0,59261701	0,65246387	0	0	0	0	0,25918452	0	0	2,24222479
2,75	0,31653776	0,72691621	0,00875392	0	0	0	0	0,72691621	0	0	1,7791241
3,29166667	0	0,25918452	0	0	0	0	0	0,59261701	0	0	0,85180153
3,83333333	0	0	0	0,29402394	0	0,14697166	0	0	0	0	0,4409956
4,375	0	0	0	0,73396959	0	0,69806148	0	0	0	0	1,43203107
4,91666667	0	0	0	0,57188435	0	0,6471204	0,51032758	0	0,51032758	0	2,23965992
5,45833333	0	0	0	0	0	0	0,74650893	0	0,74650893	0	1,49301787
6	0	0	0	0	0	0	0,38065939	0	0,38065939	0	0,76131878

Para cada t_j se obtiene la estimación de f y su representación gráfica:

$\hat{f}(t) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right)$:

t	f(t)
-0,5	0,02892224
0,04166667	0,1204242
0,58333333	0,15617214
1,125	0,0816931
1,66666667	0,1463727
2,20833333	0,26224716
2,75	0,20808362
3,29166667	0,0996254
3,83333333	0,05157817
4,375	0,16748815
4,91666667	0,26194718
5,45833333	0,17462107
6	0,08904267



Estimadores función núcleo y polinomios locales

La alternativa no paramétrica a los modelos de regresión, supone que
$Y = m(X) + e$
donde m es una función que no se supone “confinada" dentro de una familia paramétrica. Se trata de estimar m a partir de una muestra (X_1, Y_1) …; (X_n, Y_n).

Los estimadores núcleo establecen que el peso de (X_i, Y_i) en la estimación de m es

$$W_i(t, X_i) = \frac{\frac{1}{h} K\left(\frac{t - X_i}{h}\right)}{\hat{f}(t)}$$

donde K(t) es una función de densidad simétrica (por ejemplo, la normal estándar) y $\hat{f}(t)$ es un estimador kernel de la densidad como el definido en el apartado anterior.

$W_i(t, X_i)$ es, para cada i, una función de ponderación que da “mayor importancia" a los valores X_i de la variable auxiliar que están cercanos a t.

Una expresión alternativa para $W_i(t, X_i)$

$$W_i(t, X_i) = \frac{K\left(\frac{t - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t - X_i}{h}\right)}$$

A partir de los pesos W_i puede resolverse el problema de mínimos cuadrados ponderados siguiente:

$$\min_{a,b} \sum_{i=1}^{n} W_i \left(Y_i - (a + b(t - X_i))\right)^2$$

los parámetros así obtenidos dependen de t, porque los pesos W_i también dependen de t, la recta de regresión localmente ajustada alrededor de t sería :

$l_t(X) = a(t) + b(t)(t - X)$

Y la estimación de la función en el punto en donde $X = t$

$\hat{m}(t) = l_t(t) = a(t)$

Las funciones núcleo usadas en la estimación no paramétrica de la regresión son las mismas que en la densidad.

Si se generaliza al ajuste local de regresiones polinómicas de mayor grado, es decir si pretendemos estimar una forma lineal del tipo:

$\beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_q X^q$

con la salvedad de que en vez del valor X_i en la regresión lineal múltiple se utiliza el valor $(t - X_i)$. El estimador de polinomios locales de grado q asignado los pesos W_i obtenidos mediante la función núcleo se resuelve el siguiente problema de regresión polinómica ponderada:

$$\min_{\beta_0..\beta_q} \sum_{i=1}^{n} W_i \left(Y_i - \left(\beta_0 + \beta_1 (t - X_i) + ... + \beta_q (t - X_i)^q\right)\right)^2$$

Los parámetros $\hat{\beta}_j = \hat{\beta}_j(t)$ dependen del punto t en donde se realiza la estimación, y el polinomio ajustado localmente alrededor de t sería:

$$P_{q,t}(t - X) = \sum_{j=0}^{q} \hat{\beta}_j (t - X)^j$$

Siendo $m(t)$ el valor de dicho polinomio estimado en el punto en donde $X = t$:

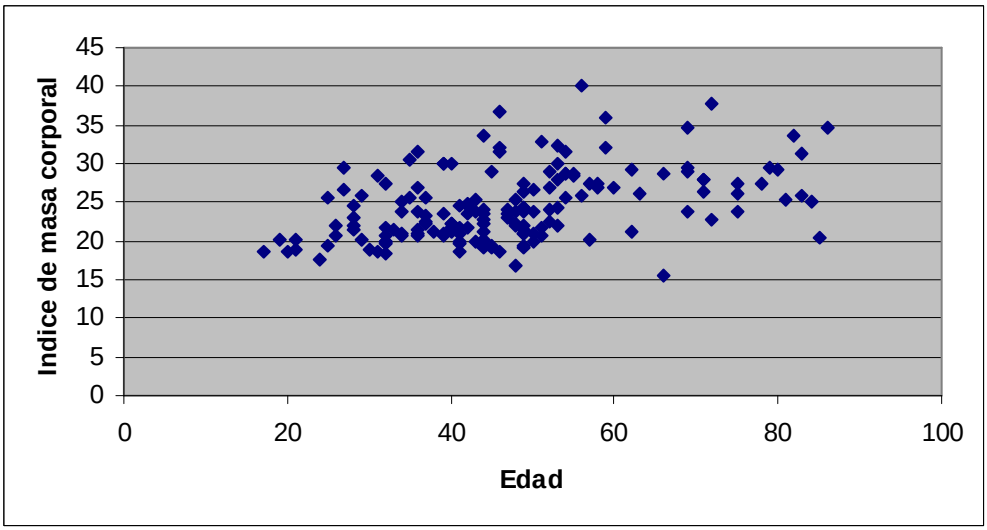
$\hat{m}_q(t) = P_{q,t}(0) = \hat{\beta}_o(t).$

En el caso particular del ajuste de un polinomio de grado cero, se obtiene el estimador de Nadaraya −Watson, o estimador núcleo de la regresión:

$$\hat{m}_K(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right)} = \sum_{i=1}^{n} W(t, X_i) Y_i$$

Ejemplo

Disponemos del siguiente conjunto de datos relativos a 163 personas con su edad y su índice de masa corporal (relación entre peso y altura):



Se va a obtener el estimador núcleo de la regresión:

$$\hat{m}_K(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right)}$$

Donde X_i es la edad de cada individuo e Y_i su masa corporal, va ha utilizarse una función núcleo de Epanechnikov, cuyo ancho de ventana sería:

$$h = (15)^{1/5} \times \left(\frac{3}{8}\right) \pi^{1/10} \times s_n \times n^{-1/5} = (15)^{1/5} \times \left(\frac{3}{8}\right) \pi^{1/10} \times 16{,}14 \times 162^{-1/5} = 4{,}22$$

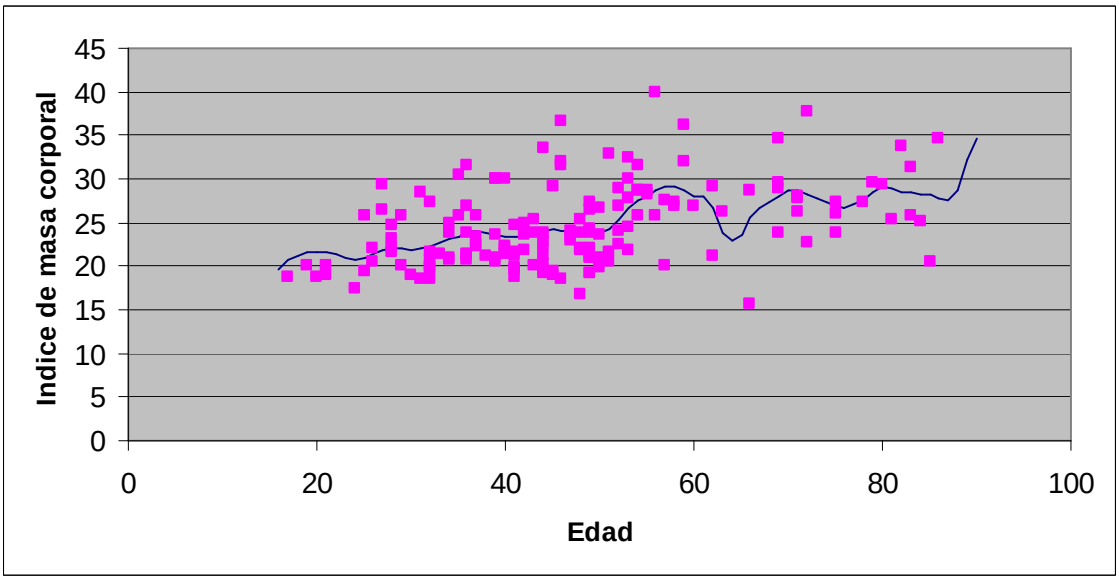
Para cada edad (t) calculamos $K\left(\frac{t - X_i}{h}\right)$:

t	$K\left(\frac{t-X_1}{h}\right)$	$K\left(\frac{t-X_2}{h}\right)$	$K\left(\frac{t-X_3}{h}\right)$	$K\left(\frac{t-X_4}{h}\right)$	$K\left(\frac{t-X_5}{h}\right)$	$K\left(\frac{t-X_i}{h}\right)$	$K\left(\frac{t-X_{159}}{h}\right)$	$K\left(\frac{t-X_{160}}{h}\right)$	$K\left(\frac{t-X_{161}}{h}\right)$	$K\left(\frac{t-X_{162}}{h}\right)$	$\sum_{i=1}^{n} K\left(\frac{t_j - X_i}{h}\right)$
16	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	1,228967175
17	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0753208	0,0000000	0,0000000	0,0000000	2,298278625
18	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,3704930	0,0000000	0,0000000	0,0000000	3,689804416
19	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,5813302	0,0000000	0,0000000	0,0000000	4,490985932
20	0,0000000	0,0753208	0,0000000	0,0000000	0,0000000	..	0,7078326	0,0000000	0,0000000	0,0000000	4,777144002
21	0,0000000	0,3704930	0,0000000	0,0000000	0,0000000	..	0,7500000	0,0000000	0,0000000	0,0000000	4,768129934
..	..	..	..	..	..	..	..	..	..	..	..
85	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	3,19280911
86	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	2,48497655
87	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	1,73497655
88	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	1,027144
89	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	0,44581379
90	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	0,07532083

Para cada edad (t) calculamos $K\left(\frac{t - X_i}{h}\right) Y_i$:

t	$K\left(\frac{t-X_1}{h}\right)Y_1$	$K\left(\frac{t-X_2}{h}\right)Y_2$	$K\left(\frac{t-X_3}{h}\right)Y_3$	$K\left(\frac{t-X_4}{h}\right)Y_4$	$K\left(\frac{t-X_5}{h}\right)Y_5$	$K\left(\frac{t-X_i}{h}\right)Y_i$	$K\left(\frac{t-X_{159}}{h}\right)Y_{159}$	$K\left(\frac{t-X_{160}}{h}\right)Y_{160}$	$K\left(\frac{t-X_{161}}{h}\right)Y_{161}$	$K\left(\frac{t-X_{162}}{h}\right)Y_{162}$	$\sum_{i=1}^{n} K\left(\frac{t-X_i}{h}\right)Y_i$
16	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	24,1149969
17	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	2,05894306	0,0000000	0,0000000	0,0000000	47,590736
18	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	10,1276624	0,0000000	0,0000000	0,0000000	78,5234969
19	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	15,8910333	0,0000000	0,0000000	0,0000000	96,7487803
20	0,0000000	1,32020961	0,0000000	0,0000000	0,0000000	..	19,3490559	0,0000000	0,0000000	0,0000000	103,586796
21	0,0000000	6,49393249	0,0000000	0,0000000	0,0000000	..	20,5017301	0,0000000	0,0000000	0,0000000	103,037148
..	..	..	..	..	..	..	..	..	..	..	..
85	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	90,1696692
86	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	69,1207607
87	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	48,0097761
88	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	29,5521528
89	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	14,3394414
90	0,0000000	0,0000000	0,0000000	0,0000000	0,0000000	..	0,0000000	0,0000000	0,0000000	0,0000000	2,60213476

En la figura siguiente se representa el estimador $\hat{m}(x)$ obtenido:



Definida la matriz

$$X_t = \begin{pmatrix} 1 & (t-X_1) & \dots & (t-X_1)^q \\ . & . & \dots & \\ . & . & \dots & \\ 1 & (t-X_n) & \dots & (t-X_n)^q \end{pmatrix}$$

Y definidos los vectores $Y = (Y_1 \dots Y_n)'$, $\varepsilon = (\varepsilon_1 \dots \varepsilon_n)'$, $\beta = (\beta_0 \dots \beta_q)'$. Se calcula la matriz de pesos W_t

$$W_t = \begin{pmatrix} W_1(X_1,t) & 0 & ... & 0 \\ 0. & W_2(X_2,t) & ... & 0 \\ . & . & ... & 0 \\ 0 & 0 & ... & W_n(X_n,t) \end{pmatrix}$$

Habría que estimar por mínimos cuadrados generalizados el modelo $Y = X\beta + \varepsilon$, cuya solución es:

$$\hat{\beta}(t) = \left(X_t' W_t X_t\right)^{-1} X_t' W_t Y$$

Pueden tomar los pesos:

$$W_i(t, X_i) = \frac{K\left(\frac{t - X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t - X_i}{h}\right)}$$

o

$$W_i(t, X_i) = K\left(\frac{t - X_i}{h}\right)$$

Ejemplo

Utilizando los datos de edades e índices de masas corporales, en libro excel "Estimador función núcleo.xls" hoja de cálculo "q2,t", se ha realizado un ejercicio para obtener un estimador de polinomio local a una función núcleo de núcleo de Epanechnikov, si se desea obtener el estimador para una edad de 65 años (t=65); la matriz X_{65} quedaría:

constante	$(65 - X_i)$	$(65 - X_i)^2$
1	-1	1
1	41	1681
1	19	361
1	20	400
1	11	121
1	-17	289
.	.	.
1	5	25
1	13	169
1	33	1089
1	34	1156
1	3	9
1	38	1444

Los pesos $W_i(65, X_i)$ serían:

$W_i(65, X_i)$
0,70783255
0
0
0
0
0
.
0
0
0
0
0
0

La matriz $X_{65}^{'} W_{65} X_{65}$ quedaría

$$X_{65}^{'} W_{65} X_{65} = \begin{bmatrix} 2,669 & -0,347 & 11,896 \\ -0,347 & 11,896 & .6,044 \\ 11,896 & -6,044 & 117,855 \end{bmatrix}$$

Y el estimador $\hat{\beta}(65) = \left(X_{65}^{'} W_{65} X_{65}\right)^{-1} X_{65}^{'} W_{65} Y$:

$$\hat{\beta}(65) = \begin{bmatrix} 22,196 \\ 0,255 \\ 0,321 \end{bmatrix}$$

Es estimador del indice de masa corporal para la edad de 65 años sería:
$\hat{m}_2(65) = \hat{\beta}_o(65) = 22,196$

Elección del parámetro de suavizado

El estimador del parámetro de suavizado h tiene una importancia crucial en el aspecto y propiedades del estimador de función de regresión. Valores pequeños de h dan mayor flexibilidad al estimador y le permiten acercarse a todos los datos observados, pero originan altos errores de predicción (sobre-estimación), valores mas altos de h ofrecerán un menor grado de ajustes a los datos pero predicican mejor, pero si h es demasiado elevado tendremos una falta de ajuste a los datos (sub-estimaciñon).

Si la cantidad de datos de que disponemos lo permite, lo habitual es obtener dos muestras una para la estimación del modelo (muestra de entrenamiento) y otra muestra para predecir (muestra de test). En este caso una medida de calidad del parametro h de suavizado es el error cuadrático medio de la población de la muestra de test:

$$ECMP_{test}(h) = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(Y_{i,t} - \hat{m}\left(X_{i,t}\right)\right)^2$$

Donde $(X_{i,t}, Y_{i,t})$, $i = 1...n_t$, es la muestra test y $\hat{m}(X)$ es el estimador no paramétrico construido con la muestra de entrenamiento. El valor h que minimice dicho error sería el parámetro de suavización elegido.

Si no de puede disponer de una muestra de test, la alternativa consiste en sacar de la muestra consecutivamente cada una de las observaciones X_i, y estimar el modelo con los restantes datos y predecir el dato ausente con el estimador obtenido, para después calcular el error de predicción. Se construye entonces la siguiente medida del error de predicción (validación cruzada) para cada h:

$$ECMP_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_i(X_i)\right)^2$$

Donde $\hat{m}_i(X)$ es el estimador obtenido al excluir la observación i-esima.
El valor h que minimice dicho error de validación cruzada sería el parámetro de suavización elegido.

Teniendo presente que el valor que predecimos $\hat{Y}_i$ no deja de ser una combinación lineal de los valores observados:

$$\hat{Y} = X\hat{\beta} = X_t\left(X_t' W_t X_t\right)^{-1} X_t' W_t Y = SY$$

Siendo $S = X_t\left(X_t' W_t X_t\right)^{-1} X_t' W_t$, matriz que se denomina de suavizado cuyo elemento (i, j) se nombra s_{ij}.
Dado que:

$$ECMP_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - s_{ii}}\right)^2$$

no es necesario ajustar las n regresiones no paramétricas, sino que vasta con evaluar todos los datos y anotar los valores de la diagonal principal de la matriz S.

Una modificación de la función anterior (Validación cruzada generalizada) permite obtener un estimador de la varianza de los errores del modelo:

$$ECMP_{GCV}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - \nu/n}\right)^2$$

Donde $\nu = Traza(S) = \sum_{i=1}^{n} s_{ii}$

Entonces:

$$ECMP_{GCV}(h) = \frac{n\hat{\sigma}_\varepsilon^2}{n - \nu}$$

Y

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n - \nu}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Bibliografía

Isabel Cañette: Estimadores de densidad basados en núcleos: cómo hallarlos en la práctica. 2004.
http://www.cmat.edu.uy/pye2008/archivos/nucleos.pdf

Pedro Delicado: Curso de Modelos no Paramétricos. Departament d'Estadística i Investigació Operativa .Universitat Politécnica de Catalunya. 14 de septiembre de 2008

http://www-eio.upc.es/~delicado/docencia/Apuntes_Models_No_Parametrics.pdf